Reliability and Bias of Different Large Language Models in IELTS Writing Score

Liangyu Wang
Anyang Normal University, Anyang, 455000, China
wly050501@163.com

ABSTRACT

Taking IELTS Writing Task 2 as the research object, this paper compares the consistency and differences between the four large language models (ChatGPT, DeepSeek, ERNIE Bot, Gemini) and human scores. The study selected 30 compositions of different levels, which were scored by two human raters and four models respectively. The data were analyzed by intra group correlation coefficient (ICC), Pearson correlation analysis and paired sample t-test.

The results showed that there was a high consistency between the two human raters (ICC=0.873). In terms of model performance, chatgpt had the highest correlation with human scores and had no significant difference; DeepSeek had a significantly low score tendency (P<0.001); ERNIE Bot also showed some deviation; Although there was no significant difference in Gemini, the correlation was relatively low. Further qualitative analysis shows that the evaluation of the large language model in high score and low score compositions is relatively consistent, while the difference is obvious in medium-level compositions.

The research shows that the large language model has certain application potential in writing assessment, but there are still differences between different models, which should be used cautiously as an auxiliary tool in practical application.

Keywords: Large Language Model; Automatic Writing Scoring; IELTS Writing; Score Consistency; Score Deviation; Qualitative Analysis.

1. INTRODUCTION

In recent years, the rapid development of large language models (LLMS) has promoted its wide application in the field of education, especially in language learning and automatic assessment [1,2]. Compared with the traditional automated essay scoring system (AES), which mainly relies on surface language features, LLMS has stronger ability in semantic understanding and context processing [3], making its application in writing evaluation become a new research hotspot.

However, it should be pointed out that writing assessment involves not only linguistic accuracy, but also complex dimensions such as argumentation structure and logical coherence [4]. Most of the existing studies focus on the performance evaluation of a single model, and there is still a lack of comparison of different large language models under the same scoring framework. In addition, there is still a lack of sufficient empirical research on whether there is systematic deviation in the model score and the consistency between the model score and human score [5].

Therefore, this study systematically analyzes the performance differences of different large language models in IELTS writing score from the perspective of multi model comparison, and evaluates them from the two dimensions of reliability and bias. This study not only helps to deepen the understanding of the application boundary of large language model in language assessment, but also provides reference for its application in actual teaching and examination scenarios.

2. LITERATURE REVIEW

2.1 Development and limitations of automatic writing scoring

As an important branch of the field of language assessment, automated essay scoring system (AES) has attracted extensive attention for a long time. Early AES system mainly scored based on surface language features, such as vocabulary use, sentence length and grammatical structure [6]. Such methods have obvious advantages in improving scoring efficiency and consistency, but their evaluation of writing quality is relatively limited, and it is difficult to effectively reflect the deep characteristics of the text, such as Argumentation Logic, discourse coherence and critical thinking ability.

With the deepening of research, scholars have gradually realized that it is difficult to fully characterize writing ability only by relying on the formal features of language [7]. Williamson et al. (2012) pointed out that although the automatic scoring system can be consistent with human raters to a certain extent, it still has obvious deviation in processing complex texts, especially in the evaluation of high-level writing ability. It can be seen that there is still a certain gap between "quantifiable features" and "real writing ability" in the traditional AES system. On this basis, the research began to introduce the neural network method to improve the scoring performance [8]. At the same time, classic automatic scoring systems such as e-rater are also important representatives in this field [9].

Overall, the early AES research provided the technical basis for automatic scoring, but its methodological limitations also pointed out the development direction for subsequent research, that is, how to more accurately simulate the human scoring process while ensuring efficiency.

2.2 Reliability in writing assessment

Writing assessment is essentially a complex cognitive activity, which requires comprehensive consideration of multiple dimensions, including task response, discourse organization, vocabulary use and grammatical accuracy [4]. Even under the standardized scoring system, there may still be differences between human raters, which makes the scoring reliability one of the core issues in the study of writing evaluation [10].

In terms of method, intra group correlation coefficient is widely used to measure the consistency between raters [11]. A high ICC value is usually regarded as a reflection of the reliability of the score. However, it should be pointed out that the reliability is not only reflected in the consistency of the scoring results, but also related to the stability and interpretability of the scoring process.

Under the background that there are still differences in manual scoring, the reliability of the scoring task is more worthy of attention when it is handed over to the automatic system. Therefore, how to achieve the same stability as human score in automatic scoring has become an important issue in current research.

2.3 Research gaps and positioning of this paper

Although the existing studies have discussed the application of automatic writing scoring and artificial intelligence in education from many angles, there are still some problems worthy of further study.

First of all, from the perspective of method development, the traditional AES system has deficiencies in deep semantic analysis [6,7]. Although LLMS has made a breakthrough in technology, the stability and consistency of its scoring performance still lack sufficient empirical support [12,13]. Secondly, from the perspective of the research object, most of the existing studies focus on a single model or theoretical analysis, and lack of systematic comparison of multiple large language models under the unified evaluation framework.

More importantly, the issue of scoring bias has not been paid enough attention, that is, whether different models have the tendency to systematically overestimate or underestimate the level of writing. This problem is not only related to the fairness of the scoring results, but also directly affects the feasibility of its application in high-risk language testing.

Based on the above shortcomings, this paper takes IELTS writing as the Research scenario, compares and analyzes several large language models under the unified scoring standard, and discusses the scoring reliability and deviation characteristics from the quantitative and qualitative levels. This study aims to provide a more empirical reference for the application of large language model in language assessment.

3. RESEARCH METHODS

3.1 Source of corpus

In this study, a total of 30 IELTS Writing Task 2 compositions were selected as the analysis corpus. The specific sources are as follows: 10 compositions were selected from IELTS blog, 10 from IELTS advantage, and another 10 compositions were publicly available real examinees' compositions.

In the process of sample selection, priority should be given to the diversity and representativeness of the text. First of all, IELTS blog and IELTS advantage, as common IELTS writing learning platforms, can better reflect the typical characteristics of writing in different grades through sorting and screening; Secondly, the real examinee's composition is

closer to the actual test situation, which helps to improve the ecological validity of the study. Therefore, the combination of the two types of corpus can help to control the quality of the text and enhance the authenticity of the data.

In addition, in the specific selection process, we should cover different writing levels as much as possible (corresponding to the IELTS 4-8 score range), so as to ensure the gradient distribution of the sample in terms of difficulty and quality. Since some of the model texts on different platforms are not clearly marked with scores, this study is mainly based on the text quality (such as language accuracy, structural integrity and argumentation level) to roughly divide them, so as to maintain the relative balance of the samples.

It should be noted that all the composition topics are not exactly the same, but they all belong to the common types of IELTS task 2 (such as opinion, discussion and problem solving). Considering that the focus of this study is on score consistency rather than topic comparison, the existence of different topics will not have a substantial impact on the research results. From the perspective of methodology, the corpus design not only ensures the availability of data, but also takes into account the internal consistency and external validity of the study.

3.2 Model selection

This study selected four representative large language models as the scoring subjects, namely, ChatGPT DeepSeek, ERNIE Bot and Gemini. The model selection is based on the following three considerations.

(1) In the field of education, relevant research has gradually focused on the application potential of large language model in automatic assessment and feedback [12]. At the same time, the review research points out that different large language models have differences in structure design, training corpus and optimization objectives, which may further affect their performance in specific tasks. Therefore, selecting several representative mainstream models for comparison is helpful to analyze their performance differences in writing scoring tasks more comprehensively.

(2) This study selects ChatGPT DeepSeek, ERNIE Bot and Gemini as the research objects, mainly based on their representativeness and wide application in the current large language model field. In addition, a variety of models have made progress in multimodal and complex language understanding tasks in recent years [2]. Previous studies have shown that the large language model represented by GPT series shows strong performance and generalization ability in a variety of language tasks [3]. ChatGPT, Gemini and other models are usually trained based on large-scale multi-source data, including Web text, books, encyclopedia knowledge and code data, so as to have strong general language understanding ability. At the same time, DeepSeek and other models introduced a high proportion of multilingual data in the training process, especially the mixed corpus of Chinese and English, which made its performance in the bilingual context representative.

(3) From the perspective of operability and accessibility, the above models can be stably called through the public platform, which can ensure the repeatability and consistency of the scoring process.

It should be noted that this study does not include other models such as Doubao and Tongyi Qianwen, which are mainly based on the following considerations: on the one hand, some models are not fully applied in academic research; On the other hand, different models still have some differences in interface stability and output consistency. In order to ensure the controllability of the study and the comparability of the results, the above four models were selected as the research object.

3.3 Data analysis method

In order to systematically evaluate the consistency and difference between the large language model and the human score, this study uses a variety of statistical methods for data analysis.

(1) Intra group correlation coefficient was used to evaluate the consistency between raters. ICC is widely used in writing scoring research, which can effectively measure the consistency of different raters' scores on the same object [11]. This study uses two-way random effect model and consistency type to ensure that the analysis results have high explanatory power.

(2) Pearson's R was used to analyze the correlation between the large language model score and the average human score. This method can reflect the strength of the linear relationship between different scoring systems, so as to evaluate the consistency of the model score in the overall trend.

(3) In order to test whether there is a systematic bias in the model score, the paired sample t-test was used to compare the mean difference between the model score and the human score. Through this method, we can judge whether the model has a significant tendency to overestimate or underestimate.

All statistical analyses were completed in jamovi software, and the significance level was set as P<.05.

As shown in Figure 1, the dataset consists of scores assigned by human raters and multiple AI models for 30 IELTS Task 2 compositions.

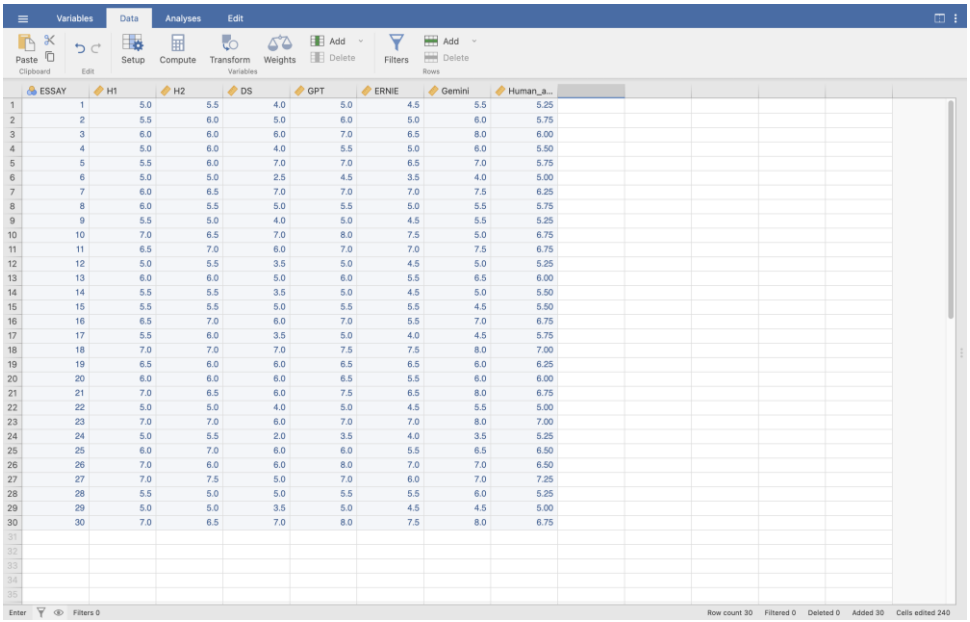

Figure 1. Raw scores of 30 IELTS Task 2 compositions.

3.3.1 Score reliability analysis (ICC)

In order to test the consistency between the two human raters, the intra group correlation coefficient was used for reliability analysis. ICC is a common indicator to measure the consistency of multiple raters' scores on the same object, which is widely used in educational evaluation research.

In this study, a two-way random effect model was used and calculated based on absolute consistency. The model assumes that both the rater and the rater are from the overall random sample, which is suitable for evaluating the overall consistency of the scoring results. ICC values range from 0 to 1. The higher the value, the stronger the score consistency. Generally speaking, ICC value higher than 0.75 can be regarded as good consistency.

Through this analysis, the reliability of manual scoring results can be verified, and reference standards can be provided for subsequent comparison.

The ICC value of the two human teachers is 0.873, which indicates that there is a high consistency between the two raters. As shown in Table 1, this supports the reliability of manual scoring as the basis for subsequent analysis.

Table 1. Intraclass Correlation Coefficient (ICC).

Model	Type	Unit	Subjects	Raters	ICC
Two way	agreement	average	30	2	0.873

Note. The analysis was performed using the irr::icc function.

3.3.2 Correlation analysis (Pearson r)

On this basis, Pearson correlation coefficient was used to analyze the correlation between the large language model score and human score. This indicator is used to measure the linear relationship between two continuous variables, and its value range is -1 to 1. The closer the value is to 1, the stronger the positive correlation between the two.

In this study, the average value of the two raters was taken as the reference for the human score, and the correlation analysis was carried out with the scores of each model. This method can reflect the consistency of different scoring subjects on the scoring trend, that is, whether the model can better capture the overall distribution characteristics of human scores.

It should be pointed out that the correlation analysis mainly focuses on the "relative consistency" of the score, that is, whether the high score is consistent with the low score, rather than the difference in the absolute value of the score.

Table 2. Correlation Matrix.

Variables	1	2	3	4	5
1.Human(average)	—				
2.DS	.753***	—			
3.ChatGPT	.851***	.919***	—		
4.ERNIE	.808***	.919***	.931***	—	
5.Gemini	.746***	.793***	.815***	.792***	—

Note. p<0.001

As shown in Table 2, the correlation between ChatGPT and human score was the strongest (r=0.851, p<0.001), followed by Ernie (r=0.808, p<0.001). The correlation between DeepSeek (r=0.753, p<0.001) and Gemini (r=0.746, p<0.001) was relatively low, but still belonged to medium to strong correlation.

In addition, a high correlation between the models was observed, especially between ChatGPT and Ernie (r=0.931), indicating that different large-scale language models tend to produce similar scoring patterns.

3.3.3 Difference test (Paired sample t-test)

To further test whether there is a systematic deviation between the large language model score and the human score, paired sample t-test was used for analysis. This method is suitable for comparing the mean difference of the same group of samples under two different conditions.

In this study, we paired the average human score of each composition with the corresponding model score, calculated the mean difference between them, and tested whether the difference was statistically significant. The significance level was set as P<0.05. When the test results are significant, it indicates that there is a systematic difference between the model score and the human score, that is, the model may have a tendency to overestimate or underestimate.

In addition, in order to understand the practical significance of the difference more comprehensively, this study also reports the mean difference and its standard error to help explain the statistical results.

Table 3. Paired Samples t-Test Results.

Comparison	t	df	p	Cohen's d
Human vs.ChatGPT	-1.26	29	.219	-0.23
Human vs. DS	4.75***	29	<.001	0.87
Human vs. ERNIE	2,53*	29	.017	0.46
Human vs. Gemini	-0.95	29	.351	-0.17

Note. p < .05*, p < .01**, p < .001***.

Table 4. Descriptive Statistics

Variables	N	M	SD
Human	30	5.97	0.69
ChatGPT	30	6.13	1.17
DS	30	5.12	1.40
ERNIE	30	5.63	1.17
Gemini	30	6.13	1.31

As shown in Table 3, ChatGPT (t (29) =-1.25, p=.219) and Gemini (t (29) =-0.95, p=.351) were not significantly different from human scores, indicating that they were highly consistent with human raters.

Descriptive statistics for all variables are presented in Table 4. As shown in Table 4, ChatGPT and Gemini had slightly higher mean scores than human raters, whereas DeepSeek tended to give lower scores.

In contrast, the score of Deepseek was significantly lower than that of human (t (29) =4.74, p<.001), indicating that DeepSeek tended to give a conservative score. Similarly, Ernie also showed a significant difference (t (29) =2.53, p=.017), although the difference was small.

3.3.4 Method integration description

It should be emphasized that the above three statistical methods are complementary in analysis. ICC is used to verify the reliability of manual scoring as the basic premise of the study; Pearson correlation analysis was used to evaluate the trend consistency between the model and human score; Paired sample t-test is used to test whether there is a significant deviation in the score.

Through the analysis of the three levels of "reliability relevance difference", we can more comprehensively evaluate the performance of the large language model in writing scoring, so as to improve the explanatory power and credibility of the research conclusion.

4. RESULT ANALYSIS

4.1 Score consistency analysis

First, the consistency between two human raters was tested. The results showed that the ICC value was 0.873, indicating that there was a high consistency between the raters. This result provides a reliable reference standard for subsequent analysis.

4.2 Correlation analysis

Pearson correlation analysis showed that there was a significant positive correlation between the scores of the major language models and human scores (P<.001). Among them, ChatGPT has the highest correlation with human score (r=0.851), indicating that it is the closest to human score in the overall score trend.

In contrast, the correlation between Gemini and human score is relatively low (r=0.746), indicating that there are still some differences in score consistency.

4.3 Difference analysis

Paired sample t-test results showed that different models had significant differences in score deviation.

Specifically, the score of DeepSeek was significantly lower than that of human (t=4.74, P<.001), indicating that the model has an obvious tendency of low score. ERNIE Bot also showed a certain degree of underestimation (t=2.53, P=.017), but the degree of deviation was relatively small.

In contrast, there was no significant difference between ChatGPT (t=-1.25, P=.219) and Gemini (t=-0.95, P=.351) and human scores, indicating that the two models were highly consistent with human scores on the scale of scores.

This result shows that there are systematic differences in the scoring standards of different large language models.

4.4 Comprehensive analysis of results

In general, there are significant differences in the performance of different large language models in IELTS writing score. ChatGPT performs well in terms of consistency and stability, which is close to the human scoring standard; Although Gemini has no significant deviation, there is still room for improvement in score consistency; ERNIE Bot and DeepSeek show a certain degree of score deviation, and DeepSeek's deviation is the most obvious.

This result shows that the correlation and difference are not completely consistent: some models may be close to human scores in the scoring trend, but there are still systematic deviations in the specific scores; On the contrary, it may be close to the human score on the average, but the score distribution is not stable enough.

Therefore, it is difficult to comprehensively evaluate the performance of the model only by relying on a single indicator, and it is necessary to conduct a comprehensive analysis with a variety of statistical methods.

These results provide an important basis for the follow-up discussion on the scoring mechanism of the model and its application limitations.

4.5 Qualitative analysis: comparison of model evaluation characteristics

In order to further understand the differences between different large language models in the scoring process, this study makes a qualitative analysis of the brief evaluation texts generated by each model. By summing up the feedback of compositions in different score bands, the following characteristics can be summarized.

The results of qualitative analysis show that in high score compositions, each model can generally recognize the advantages of clear structure and sufficient argument, but there are differences in the evaluation focus. In low score compositions, the models show high consistency in identifying grammatical errors and logical problems. In contrast, in the medium-level writing, the differences between different models in the evaluation focus and scoring criteria are more obvious, indicating that such texts are more challenging to the automatic scoring system.

In addition, from the perspective of evaluation style, different models also have differences in feedback expression. For example, some models are simple and focus on key issues; Some models tend to provide more detailed explanations. This style difference may have an impact on the scoring results. From the perspective of scoring tendency, some models show a relatively strict or loose tendency in evaluation, which is consistent with the score deviation observed in quantitative analysis. For example, models with low scores tend to emphasize problems more frequently in the evaluation, while models with relatively close scores to humans maintain a relative balance between advantages and disadvantages.

Qualitative analysis shows that there are not only score differences in writing evaluation, but also systematic differences in evaluation focus and expression methods. These findings are helpful to further understand the results of quantitative analysis and provide support for subsequent discussion.

5. DISCUSSION

The results of this study to some extent support the current view that the large language model has potential application in writing assessment [12]. However, it should be emphasized that this potential is not consistent across all models. The differences between different models in score consistency and score deviation indicate that their internal mechanism and training data may have an important impact on the score results [13].

In addition, the results of qualitative analysis further show that the performance of the model in dealing with different levels of writing is not balanced. Compared with high score and low score compositions, medium level texts tend to contain more complex linguistic and structural features, which may increase the uncertainty of model judgment. This finding suggests that in practical application, LLMS may be more suitable as an auxiliary evaluation tool rather than a complete substitute for human raters.

6. CONCLUSION

Based on 30 IELTS Writing Task 2 compositions, this study systematically compares the performance of four major language models in writing scoring. Through the combination of quantitative analysis and qualitative analysis, the study found that there were significant differences in score consistency and score deviation between different models.

Specifically, ChatGPT shows high consistency in both score trend and numerical level, and is the closest model to human score among the four models; Gemini did not show significant score deviation, but there was still room for improvement in score stability; ERNIE Bot and DeepSeek show different degrees of systematic deviation, and DeepSeek's deviation is the most obvious.

On the whole, the large language model shows certain application potential in writing assessment, especially in improving scoring efficiency and providing auxiliary feedback. However, at present, the scoring results are still difficult to completely replace the manual scoring. In practical application, it should be used as an auxiliary tool and combined with manual scoring to ensure the accuracy and fairness of the evaluation.

From the perspective of multi model comparison, this study provides an empirical basis for understanding the performance of large language model in language assessment, and also provides a reference direction for future related research, which has a certain reference significance for understanding the application boundary of large language model in high-risk language testing.

REFERENCES

[1] Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. IEEE Access, 8, 75264-75278. https://doi.org/10.1109/ACCESS.2020.2988510

[2] Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. Computers and Education: Artificial Intelligence, 1, 100001. https://doi.org/10.1016/j.caeai.2020.100001

[3] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Amodei, D., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[4] Weigle, S. C. (2002). Assessing writing. Cambridge University Press.

[5] Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International. https://doi.org/10.1080/14703297.2023.2190148

[6] Dikli, S. (2006). An overview of automated scoring of essays. The Journal of Technology, Learning and Assessment, 5(1), 1-35.

[7] Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. Educational Measurement: Issues and Practice, 31(1), 2-13.

[8] Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1882-1891).

[9] Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. The Journal of Technology, Learning and Assessment, 4(3), 1-31.

[10] Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. Language Assessment Quarterly, 9(3), 270-292.

[11] Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

[12] Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

[13] Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learning Environments, 10(1), 15. https://doi.org/10.1186/s40561-023-00237-x